CHINESE WORD SEGMENTATION AND ITS EFFECTS ON CHINESE INFORMATION RETRIEVAL

by
Li Wen

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina
April, 2003

Approved by:

_____

Advisor

Li Wen. Chinese Word Segmentation And its Effects on Chinese Information Retrieval. A Master's paper for the M.S. in I.S. degree. April, 2003. 46 pages. Advisor: Gregory B. Newby

This experiment tests the effectiveness of Chinese information retrieval using a segmenter that is developed with dictionary-based Maximum Forward Matching algorithm. IRTOOLS, an IR system developed at UNC Chapel Hill, is used as the platform.

This study finds that less accurate segmentation will not necessarily yield worse information retrieval results. As a matter of fact, allowing two-character words only in the dictionary produced the best retrieval results in terms of precision and recall. Allowing longer words in the dictionary will lead to the missing of index words -- the problem of over-specification. However, long-word indexing can produce better results when the long-word is also used in queries.

Headings:

Information Retrieval -- Chinese

Word Segmentation -- Chinese

TABLE OF CONTENTS

1. Introduction

It is a well-known fact that Chinese sentences, unlike its English counterparts, are written continuously. There are no white spaces that separate each “word”. A single Chinese character, or hanzi, is for most of the time, both a syllable and a morpheme. Even though a single character does bear some meaning, words that bear more complex meaning are often created by combining more than one morpheme together. Despite of the fact that words are written continuously without any delimiters, Chinese people can agree on most of the boundaries of words according to context information.

However, the definition of Chinese words is inherently ambiguous. A compound in English can be easily identified since it contains more than one space-separated words, but it is hard to distinguish a compound phase from a simple word in Chinese. For example, 新华社（Xinhua New Agency）may either be identified as a simple word or as a compound, that is 新华 (Xinhua) + 社(organized body). In other cases, a derived word is normally a single word in English, but it may be identified as a compound in Chinese. For instance, pianist (钢琴家) is derived from piano to represent the kind of person who plays piano. In Chinese the same meaning is formed by combining 钢琴 (piano) and 家 (a specialist in certain field). To either regard pianist as two words or one word is both legitimate. In empirical studies (Sprout et al., 1996; Wu and Fung, 1994; Hoosain, 1991; Tsai, McConkie, and Zheng 1998) where subjects are asked to mark the

boundary of words in Chinese sentences, researchers found that Chinese people don't always agree with each other on where the word boundary should be. An agreement rate as low as 75% is common. Miller, Chen, and Zhang (2000) have further demonstrated that English speakers showed strong agreement on word boundaries when asked to parse sentences (whose spaces between words were removed) into words. Chinese speakers, on the other hand, showed considerable disagreement.

The lack of word delimiters, and the lack of a clear sense of word among native speakers impose new problems for information retrieval in Chinese than in English.

For English language, information retrieval generally involves indexing and ranking. Indexing is a process of representing what a document is about. Words, word frequency (how many time a word occurs in a document), distance between two words, and the location of a word in a document are common candidates for representing a document. The quality of an information retrieval system is most commonly measured by precision and/or recall, depending on the functionality of the system. Precision is a measure of correctness by calculating the proportion of relevant documents from all retrieved documents. Recall is a measure of completeness by calculating the proportion of relevant documents from all relevant documents in the collection. Google.com, for example, would emphasize on precision, since its aim is not to retrieval all the relevant documents a user wants. In fact, as long as the first few documents are good enough, its job is done.

A legal system on the other hand might put more efforts on recall, to make sure that documents about a particular case are completely retrieved out. The differences in precision and recall are produced with different ranking methods. Why a document ranks higher, thus gets retrieved earlier, than another document depends on how similar they are with the query. A simple way of calculating the similarity between a document and a query is to regard both of them as a vector of words, and to calculate their distance mathematically. Each word can further be weighted according to its discrimination power (the word *the* appears so frequently in documents has lower discrimination power than the word *uranium*, thus plays a less important role in representing the document), word frequency in one document and word frequency in all documents. This weighting method is also called TF-IDF, and is widely used. Some words occur so often, such as the, a, of etc, that they carry no useful information on the document, and are thus not indexed. They are generally called stop words.

For Chinese language, a simple way for indexing is to regard each character as a word, and index by character. As mentioned above, discrimination power, word frequency in a document and the number of documents in which the term occurs are important for a word to effectively represent a document. However frequently-used-Chinese-characters are not discriminative enough. There are approximately 2,500 most-frequently-used characters and 1,000 second most-frequently-used characters according to the Xiandai Hanyu Changyongzi Biao (Modern Chinese Commonly-Used Word List), compiled by

the national language committee and national education committee in 1987. According to dictionary, Hanyu Dazidian, published by Hubei publishers of Sichuan province in 1986, there are over fifty-six thousand characters. This would result in an index that is highly unevenly distributed when most document information will cluster around the 3,500 also characters. On the other hand, when those frequently used characters are combined with each other, they can have completely different meanings. For example, for the same character "海", it can be used to form 上海 (Shanghai) or 海军 (navy) or 海关 (custom) and many more. Therefore, even in approaches that used single-character based method (Sproat & Shih, 1990), other algorithm is generally used to find out the positional relationship among characters.

The structure of this article is as follows: in first part, it addresses the question of why word identification is important in information retrieval; in second part, it gives a literature view of different approaches on Chinese segmentation; in third part, it describes the experiment of Chinese segmenter on IRTOOLS -- the information retrieval system developed at UNC Chapel Hill; in fourth part, it describes the information retrieval results and finally talks about the conclusion and future work based on the study.

2. The Importance of Word Identification in Information Retrieval

As discussed above, an important factor that determines the ranking of a document for a

particular query is the similarity between the document and query. Although there are different algorithms used to measure the similarity (vector space model, latent semantic indexing), a document that contains all the query terms, query terms are close to each other in the document, with a high frequency of occurrences are generally more similar to a document that either don't contain all the query terms, or terms are far apart from each other, or with fewer occurrences.

A word can provide similarity information in three ways: the distance among characters or letters, the sequence they appear, and the probability they appear together. Assuming that English texts didn't have spaces between each words, and are indexed by alphabetic letters. The likelihood that letters in queries would appear in a document would be significantly increased. As a result, an information retrieval system would find it very difficult to discriminate one document from another against the query. Even though the IR system can incorporate statistical method to calculate the co-occurrence of each letter with other letters, and the most frequent occurring sequences of letters etc, there are at least 26^{10} kinds of combinations, assuming most English words are shorter than 10 letters. It would require large amount of computing power to calculate their relationship with each other and the probability an instance will occur. On the other hand, when English documents are indexed by words, the IR system only need to deal with less than one million entries. Webster's Third New International Dictionary, for example, contains over 450,000 vocabulary entries.

Chinese characters are better candidates than English letters due to the following two reasons: Chinese characters are far more discriminative than English letters and second, most Chinese words are no more than two characters long. Even if we assume that there are 5,000 characters that would normally occur in Chinese newspapers, a Chinese IR system will only need to calculate the relationship and distance among 25 millions possible combinations. Although it is more than 25 times the size of English word index, it is still achievable, combined with other searching algorithms, such as the weighted finite-state transducer used by Sproat et al (1996). This helps to explain why pure statistical methods can produce decent results using character based indexing as in Sproat et al's study in 1990 and 1996.

Despite of the fact that Chinese characters are self-discriminative to some degree, indexing by words can add similarity information, thus greatly reducing the computing resources that are needed to compute such kind of relationship. In this sense, dictionary approach is preferred to non-dictionary approach, and should produce a result no worse than character-based approach.

3. Previous Work on Chinese Segmentation

Chinese text segmentation is a widely researched area because of the complexity and its importance to machine translation, natural language processing and information retrieval.

There are a variety of approaches researchers have used and are categorized differently by differently researchers. For example, Khoo et all (2002) summarized previous work into four types: statistical method, dictionary-based methods, syntax-based methods and conceptual methods. Foo & Li (2002) instead categories them into two large approaches: character-based and word-based approaches. Under character-based approach, there is single-character based approach and multi-character based approach. For word-based approaches, there are statistic-based, dictionary-based and hybrid approaches. Dictionary-based approaches can be further divided into phrase approach and component approach and so on. Wu & Tseng (1993) divided all the approaches into two categories: single character-based approaches and multicharacter-based approaches. Under each category, there are dictionary and non-dictionary based approaches. Dictionary based multi-character approach is further divided according to phrase/word and linguistic/non-linguistic. Contrary to Foo & Li' and Wu & Tseng's hierarchical structure of classification, I am more leaning toward to a flatter one, since as mentioned before, word boundary is very vague in Chinese language, and 1-gram/single character, bi-gram, tri-gram, statistics, dictionaries and heuristic rules are simply techniques that are normally combined to produce better results. Therefore I would rather treat all of them as techniques that can be used for Chinese text segmentation.

3.1. Single-character technique

Single character technique treats each Chinese character as an index element. This technique can be very easily realized since Chinese characters are encoded with two bytes of data in machine code. Furthermore, encoding techniques, such as Unicode, UTF-8 allows a programmer to get a Chinese character out of incoming stream the same way as getting an English letter. Some researchers (Huang & Robertson, 1997; Nie, Chevallet, & Bruandet, 1997; Smeaton & Wilkinson, 1996; Buckley, Singhal, & Mitra, 1996) have reported comparative results by only using single-character technique with approaches using multi-character techniques.

The reason why single-character technique can work in IR system might be that Chinese characters are quite discriminative in meaning, if not as discriminative as words or phrases. Furthermore, it is possible to make up the distance, sequence and co-occurrence information among characters with a powerful algorithm.

3.2. Dictionary

Dictionary is one of the most frequently used techniques in Chinese segmentation. There are a variety of ways in which a dictionary can be used. Two of the most commonly used are Maximum Backward Matching (MBM) and Maximum Forward Matching (MFM). In MFM, the match starts from the left of a Chinese sentence. A number of characters, from the longest to the shortest, are extracted from the sentence to match against the words in

the dictionary, until a match is found. When a match is found, the starting point is moved forward to the end of the word found, and start another round of greedy matching. The rational for this algorithm is that longer words are more discriminative than shorter words therefore should be extracted first. In example 1, if 中国, 外经贸部, 外, 经贸, 部 (China, Ministry of Foreign Trade and Economic Cooperation, outside, trade and economy, department) are all in the dictionary, then it will be segmented as 中国 外经贸部 rather than 中国 外 经贸 部. On the contrary, Maximum Backward Matching starts from the end of the sentence.

Ex 1: 中国外经贸部在北京设立反倾销公开信息查阅室

Ministry of Foreign Trade and Economic Cooperation of China set up anti-dumping public information consulting office

This technique has been widely used and approved to be quite effective. (Chen & Liu, 1992; Cheng, et al, 1999; Leung et al, 1996; Li et al, 1991; Yao el al, 1990; Yeh & Lee, 1991).

There are many problems with pure dictionary approach. As we can notice that the dictionary used in the segmenter basically determines the universe of terms that will be indexed. If a word does not appear in the dictionary, there is no way the segmenter will extract it from a document. This problem is especially crucial to information retrieval,

since named entities, such as names of persons or locations, are often used as query terms. And because new names occur so quickly, it is hard, if not possible for dictionaries to keep up to date at that speed. Many researchers have combined statistical and/or heuristic methods to find out named identities (Lee et al, 1999).

Another problem dictionary-based approach needs to solve is segmentation ambiguity. There are two kinds of ambiguities: crossed ambiguity and combined ambiguity (Yu & Yu). Crossed ambiguity occurs when a character can be combined with both previous and subsequent character to form a word. In example 2, 信 (letter) can be combined with 公开 or form 公开信 (public letter), which is a commonly used phrase. It can also be combined with 息 to form 信息 (information). If 公开信, 公开, 信息, 信 are all in the dictionary, by using forward greedy matching, it will be segmented to 公开信 息（public letter + breath）. However, if backward greedy matching is used, the four characters will be segmented into 公开 信息 (public + information). A Chinese speaker can easily find out that the second segmentation is correct, but it will be very difficulty for a computer to make judgment on contextual meanings. Combined ambiguity happens when words can also be combined to form new words. In example 3, all of them, 希望工程 (The Hope Project), 希望 (hope), 工程 (project) are legitimate words. It is correct for the segmenter to take 希望工程 as a whole or segment it into 希望 and 工程. A information retrieval system might do a better job to index such documents under 希望工程 (The Hope Project), when this phrase has become a well-accepted phrase that

embodies the movements of Chinese government to solve the education problems of children of poor families. Under other circumstances, over precision may be counter-effective for information retrieval systems. As a matter of fact, combined ambiguity occurs with other languages as well. For example, should a document containing University of North Carolina be indexed by University of North Carolina, or University + North Carolina or University + North + Carolina? If it is indexed too precisely (indexed by University of North Carolina), a query formed by "North Carolina education" might miss it. The low degree of agreements among native speakers when segmentation is concerned suggests that even human beings can't solve the combined ambiguity quite well.

Ex 2: cross ambiguity

公开信息 -> 公开信 息 （public letter + breath）

公开信息 -> 公开 信息 (public + information)

Ex 3: combined ambiguity

希望工程 (The Hope Project)

希望 工程 （hope + project）

3.3. n-gram technique

In n-gram approach, the most often used are 1-gram, bi-gram and tri-gram. It is based on two observations. First is the high occurrence of short-word (less than three characters) in Chinese language. According to Liu's (1987) study, it is estimated that 5% of all words are one-character words, 75% are two character words, 14% are three-character words, and 6% have more than three characters. Therefore about 94% of the words in documents are short words. Another study by He (1983) also found that 93.2% of the words in Chinese documents are short words. Another observation is that words that are longer than two characters can be formed by a combination of one-character or two-character words (Wang, 1985; Wang & Xiao, 1984).

Bi-gram is a technique that divides a Chinese sentence from left to right into overlapping two-character combinations. In example 1, 中国外经贸部在北京设立反倾销公开信息查阅室, using by-gram technique, becomes 中国 国外 外经 经贸 贸部 部在 在北 北京 京设 设立 立反 反倾 倾销 销公 公开 开信 信息 息查 查阅 阅室. One problem with bi-gram technique is the significant increase of number of indexes in a information retrieval system. If the same sentence is segmented by human being, one likely way is 中国 外经贸部 在 北京 设立 反倾销 公开 信息 查阅室. Even though some ambiguity will occur among different people, the number of indexes should be much smaller than that produced by bi-gram technique. On the other hand, bi-gram handles the problem of unknown words better than dictionary-based approach. N-gram techniques are seldom used alone. They are most often combined with statistical methods

to eliminate string of n-grams that are not correct words (Yang et al, 2000; Khoo et al, 2002).

3.4. Statistical Methods

Statistics technique is widely used in Chinese segmentation, if not the most widely used. This technique is based on the observation that meanings in Chinese are based on words, and association of character(s) in a word should be significantly higher than non-words. There are a variety of statistical methods researchers have tried with Chinese segmentation. One of the most common is mutual information formula.

In Yang et al's (2000) study, mutual information is used to determine the extraction of n-grams from a sentence. They defined mutual information as the "statistical measurement of association between two events, a and b." Mutual information of ci and cj is computed by I (ci, cj) = log2(Nf[ci, cj]/f[ci]f[cj]), where N denotes the total number of characters in the collection, f[ci, cj] denotes the frequency of ci followed by cj, and f[ci] denotes the frequency of ci. Sentence is first segmented into bi-grams, and those bi-grams that have high mutual information are extracted first. Characters that are not extracted are segmented into 1-gram. Then tri-grams are formed by combining adjacent 1-gram and bi-gram. Mutual information is computed again to extract tri-grams that have high mutual information value. Finally, quadra-grams are formed by combining adjacent 1-, bi- and

tri-grams. Candidate 4-characters words are extracted when the mutual information if high.

Khoo et al (2002) used contextual information formulas, rather than mutual information formula to calculate the closeness of relationship for bi-grams and tri-grams. Unlike mutual information formula, which is only concerned about term frequency in the collection, contextual information is calculated using both term frequency and weighted document frequency to determine a word boundary. Document frequency is defined as the number of documents in the collection containing n-grams. Weighted document frequency is calculated by taking into account the number of times the n-gram concerned occurs in each document. They found that contextual information formula performs substantially better than the mutual information formula.

Others include stochastic finite state model (Sproat et al, 1996), hidden Markov model (Allen, 1995). Interested readers can refer to Khoo et al's (2002) paper for a detailed overview of studies using statistical methods.

3.5. Linguistic Rules

Many researchers use linguistic knowledge as ad-hoc rules to identify word boundary, including morphology and grammar.

Wu & Tseng (1995) first segment texts into two character words using backward matching against a dictionary. All words, 1-gram and bi-gram, are tagged with syntactic categories, such as verbs, nouns, adjectives etc. The segments are then parsed based on grammar to form more complex words or phrases. Their study is inspired by English phrase-analysis-based text retrieval.

Kwok (1999) uses a very small dictionary initially to segment Chinese text using maximum forward matching algorithm. Linguistic rules are applied to unsegmented chunks of characters for further segmentation. Such rules include rule d, rule 2, rule 3 and rule e. Rule d regards two adjacent identical characters as a word; Rule 2 deals with quantifiers, since quantifiers in Chinese are composed of numeric and measurement. Rule e uses bi-gram technique to segment the remain chunk further.

Sproat & Chang's (1996) study uses linguistic knowledge to deal with morphologically derived words, personal names and transliterated foreign names based on a stochastic finite-state model. One interesting thing about their method in identifying Chinese personal names it to get a list of characters that would occur in family name and given name, with frequency of occurrence in a name, and them estimate the probability of a potential name as the product of the probability of finding any name in text and the probabilities of all characters in either family name or given name. They have a list of characters that are particularly common in transliterations as well to identify

transliterated foreign names.

3.6. Effectiveness of Word Segmentation and Its Effects on Information Retrieval

Traditional, there are three algorithms that can be used to judge the effectiveness of segmentation results (Palmer & Burger, 1997). One is Binary Decision (BD), which is the percentage of number of correct boundary judgments over the total number of characters. For example, a Chinese sentence ABCDEFG has seven characters. Segmenting a Chinese sentence is to identify where the word boundary is. If AB CD E FG is the correct segmentation, boundary decision has to be made after B, D, E, and G. If after segmentation, the sentence becomes AB C D EFG, the BD percentage is 3/7 or approximately 0.43, since B, D and G are correctly identified. The problem with this calculation is that it can only discover missed binary decision, such as E in our example, but can't penalize added boundary decisions, such as C in our case. Therefore, a segmentation result of A B C D E F G will have the same BD score as correctly segmented sentence.

The second is Boundary Recall/Precision (BRP). The idea of precision and recall is borrowed from the notion of precision/recall in judging information retrieval results. Recall (R) is defined as the percentage of correct boundaries identified, that is the BD percentage; precision (P) is defined as the percentage of correctly identified boundaries

over the total identified boundaries. In our example above, if the segmentation result is AB C D EFG, the R-value is the same as BD score, that is 0.43. For P value, the total identified boundaries are B, C, D, and G, and the correctly identified boundaries are B, D, G. Therefore, the P value is 3/4 or 0.75. R and P values are then used to calculate balanced F- measure using the following equation (Rijsbergen, 1979): F=2PR/(P+R). Our F value becomes 0.5 after replacing the letters with values. If the segmentation turns out to be A B C D E F G, we can easily find out that F= R = P = 0.43. Therefore it is a worse segmentation result than that of AB C D EFG (0.5).

A third measure is Word Recall/Precision (WPR). Here instead of judging effectiveness of segmentation according to correctly identified boundaries, we are now more concerned about the percentage of words in the manually segmented text identified by the segmenter (Recall) and the percentage of words identified by the segmenter that are also identified manually. In our example, four words are identified manually: AB CD E FG. Four words are identified by segmenter: AB C D EFG. There is only one word correctly identified, AB, therefore R-value is 1/4 or 0.25. P-value is also 0.25. F-value becomes 0.25. If the segmentation result is A B C D E F G, no words are correctly identified, therefore F = R = P=0.

The above three scoring methods all assume a manual segmentation result, that is perfectly segmented. One problem of this kind of measure is that, as we have discussed at

the beginning of this paper, the definition of words is vague in Chinese, and word boundary agreement is as low as 75%. How can a researcher be sure that a segmentation method is better than the other based on ambiguous reference?

Another approach for evaluating the effectiveness of segmentation is using information retrieval results by calculating the precision and recall. The common wisdom is that better segmentation will lead to better information retrieval results. Interestingly, researchers (Foo & Li, 2002) found out that correct segmentation may not be so important after all. In their efforts to find out to what extend does correct segmentation affect information retrieval, Foo & Li (2002) noticed that manual segmentation does not always work better than character-based segmentation. The existence of long-words (more than two characters) may have adverse effects on information retrieval. They failed to find any evident that ambiguous words will significantly affect IR. However they did find that the correct identification of two character words could significant improve the performance of IR system. Kwok (1999) also found that good retrieval was not dependent on accurate word segmentation; approximate segmentation into short-words would do. Short words, or simple words mean the smallest independent unit of a sentence that has meaning on its own (Khoo, 2002), as compared with compound words or phrases. The latter is generally called long words. Since almost 93% of the Chinese words are less than 4-character long (He, 1983; Liu, 1987), many experiments are done using words that have no more than 4 characters.

4. The Experiment

In our experiment, IRTOOLS, an information retrieval system developed at University of North Carolina Chapel Hill is used for the testing. IRTOOLS is developed to serve as a platform, where new IR techniques can be tested against terabyte-scale collections. It also serves as a basis for annual TREC entries. One of IRTOOLS's major tasks is to handle cross-language information retrieval request. Currently it is able to index and retrieve Arabic and English languages with the same tokenizer. However new segmenter has to be developed in order to handle Chinese documents.

Maximum Forward Match (MFM) is chosen for our experiment because of the simplicity, speed and effectiveness. The dictionary we used is created by combining two word lists found on the web. One is a Chinese to English word list from the Linguistic Data Consortium and the other is a word list called *duoyuanpinyin ciku* for richwin found at http://www.geocities.com/hao510/wordlist/. The former has about 24,000 entries and the latter contains 120,300 entries of Chinese words.

The test data used are TREC-5 Chinese collection. There are 24,977 error free documents in total, 38 MB in size, and were collected from Xinhua News Agency, one of the most influential state news agencies. Here is how TREC test data can be used. First all text documents are reformatted in XML. Each physical file, such as a word file that has a name to it, contains multiple documents. Each document starts with a document tag (refer

to Appendix A) -- <DOC>, and ends with a closing tag -- </DOC>. Each document has a document ID and a document number. TREC also provide standard queries to the data collection. There are 28 queries for TREC-5 collection. Experts made relevance judgments for each query-document pair. Therefore, after running this standard set of queries against the standard collection of documents, software will be able to automatically calculate the precision and recall. The software to be used is trec_eval, which can be freely downloaded from ftp://ftp.cs.cornell.edu/pub/smart/.

Our purpose is first to find out how the length of words in the dictionary will affect information retrieval. Second, how our system is doing in terms of precision and recall using a simple MFM algorithm, and discover ways that can be used to improve the retrieval performance on our system using Chinese documents.

Theoretically, one can expect to get better retrieval results when both query and documents are segmented using the same segmenter. This assumption has been empirically approved by Foo & Li (2002). The lengths of words in our dictionary ranges from 1 to 22. We plan to use n-characters words to segment our data collection, where n is 2, less than or equal to 3, less than or equal to 4, less than or equal to 5 and any lengths. Furthermore, query and documents are segmented using the same segmenter. In this way, we can find out how the length of words in our dictionary can affect the effectiveness of information retrieval in terms of precision and recall.

5. Results

I ran five tests with different n-character words where n=2, n<=3, n<=4, n<=5 and n=all possible lengths respectively. During each run, all queries are first segmented with the allowed word length(s) in the dictionary. Then documents are segmented and indexed using IRTOOLS. Finally, all queries are tested against the generated indexes. There are originally 28 queries for TREC-5 Chinese collection. However two of them don't have any relevant documents in the collection. Therefore, we have 26 queries in total (refer to Appendix B). TREC queries are generally longer than what might be used in everyday search on google.com. For example, one of the queries is: 苏联 在 海湾 战争 中 如何 担任 调停 的 角色 (How does Soviet Union carry out the role of mediator in the Gulf War).

After segmentation and indexing, 24,977 documents are successfully processed in our system. Some of them failed, since our current system can't handle a document that contains illegal coding, and therefore can't be converted into wide characters in our system. I plan to improve this with more tolerant code. There are 804 documents that are relevant to at least one of the queries.

Table 1 and Graph 1 shows each run's precision values at certain recall levels. Their average precisions are recorded as precision2, precision 23, precision234, precision2345,

and precisionAll according to the length(s) of words allowed in the segmenter. Table 2 and Graph 2 shows the precision values in the first 5/10/15/20/30 documents.

Table 1: Recall-Precision for five runs

Recall	Precision2	Precision23	Precision234	Precision2345	PrecisionAll
0.00	0.7036	0.6404	0.6609	0.6594	0.6288
0.10	0.5504	0.5522	0.5222	0.5218	0.5206
0.20	0.4625	0.4367	0.4340	0.4332	0.4328
0.30	0.4152	0.4029	0.4056	0.4054	0.4037
0.40	0.3819	0.3613	0.3585	0.3588	0.3553
0.50	0.3590	0.3235	0.3262	0.3258	0.3235
0.60	0.3000	0.2775	0.2833	0.2839	0.2825
0.70	0.2628	0.2521	0.2431	0.2433	0.2408
0.80	0.1962	0.1762	0.1705	0.1660	0.1643
0.90	0.1249	0.1244	0.1222	0.1209	0.1191
1.00	0.0789	0.0758	0.0755	0.0755	0.0755

Graph 1: Recall-Precision Curves

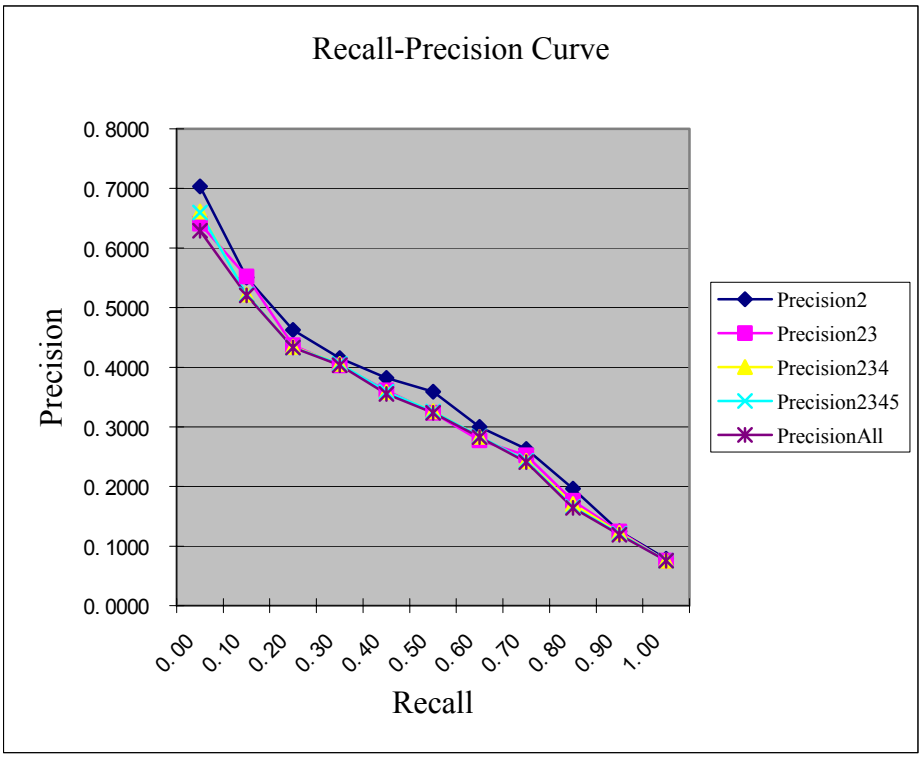
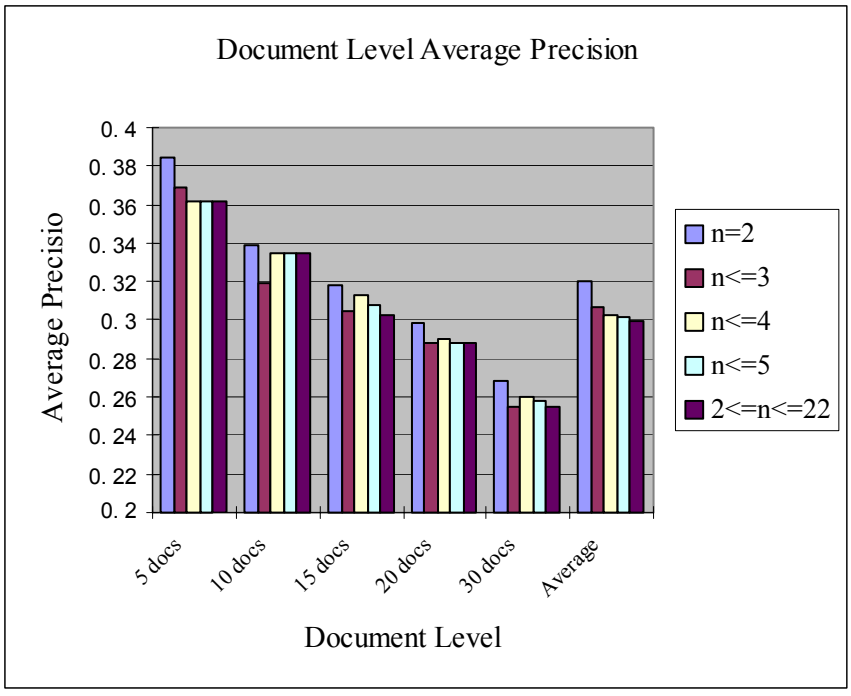


Table 2: Percentage variation of document level precision where 2-character words segmentation is the base level

Precision	n=2	base	n<=3	Increase/ Decrease (%)	n<=4	Increase/ Decrease (%)
5 docs	0.3846	1	0.3692	-4.00%	0.3615	-6.01%
10 docs	0.3385	1	0.3192	-5.70%	0.3346	.15%
15 docs	0.3179	1	0.3051	-4.03%	0.3128	.60%
20 docs	0.2981	1	0.2885	-3.22%	0.2904	-2.58%
30 docs	0.2679	1	0.2551	-4.78%	0.2603	-2.84%
Average	0.3204	1	0.3064	-4.37%	0.3025	-5.59%

Precision	n=2	base	n<=5	Increase/ Decrease (%)	All lengths	Increase/ Decrease (%)
5 docs	0.3846	1	0.3615	-6.01%	0.3615	-6.01%
10 docs	0.3385	1	0.3346	.15%	0.3346	.15%
15 docs	0.3179	1	0.3077	-3.21%	0.3026	-4.81%
20 docs	0.2981	1	0.2885	-3.22%	0.2885	-3.22%
30 docs	0.2679	1	0.2577	-3.81%	0.2551	-4.78%
Average	0.3204	1	0.3017	-5.84%	0.2995	-6.52%

Graph 2: Document Level Average Precisions



From recall-precision curves, the system gets its highest performance when only

2-character words are used in the segmenter. When both 2-character and 3-character words are allowed, performance fluctuates, and sometimes can be as good as or even slightly better than that of 2-character word only. The worst performances are observed when words of all lengths are entered into the dictionary. We are also concerned about the precision values in the first 30 ranked documents. As shown in Graph 2, approach using two-character words in the segmenter shows higher precisions consistently at all document levels.

A close look at the query segmentation results (Appendix C) using different approaches, we found that segmenter (n=2) doesn't produce the best segmentation results, if not the worst. On the other hand, segmenter with all possible length words in our dictionary produced a result that is more plausible for native speakers. For instance, it correctly identified many common phrases and long-words: 联合国(United Nations), 核电站 (nuclear power plant), 世界贸易组织 (World Trade Organization), 边境贸易 (border trade), 穆斯林(Muslin), 中国大陆 (Mainland China), 知识产权(intellectual rights) etc. Segmenter with two-character words failed to identify important words, such as 伊拉克 (Iraq), 爱滋病 (AIDS), 穆斯林(Muslin) etc.

To further understand why less effective segmentation (2-character only) produced better retrieval results, in terms of precision, I draw the following chart and graph (table 3): a query by query comparison among 2-character segmentation, 2/3-character and

all-character segmentation.

Table 3: Query by Query Comparison among 2-character, 2/3-character and all-character segmentation

Queryid	n=2	N<=3	increase/ decrease	all lengths	increase/ decrease
2	0.1084	0.1062	-2.03%	0.0279	-74.26%
3	0.1311	0.0730	-44.32%	0.0630	-51.95%
4	0.2026	0.1410	-30.40%	0.1521	-24.93%
5	0.0811	0.0825	1.73%	0.0531	-34.53%
6	0.0822	0.0915	11.31%	0.2874	249.64%
7	0.1076	0.1078	0.19%	0.1009	-6.23%
8	0.3076	0.3091	0.49%	0.3209	4.32%
9	0.2653	0.2933	10.55%	0.3103	16.96%
10	0.3554	0.3392	-4.56%	0.2425	-31.77%
11	0.3837	0.3546	-7.58%	0.3350	2.69%
12	0.1672	0.1680	0.48%	0.1504	0.05%
13	0.0000	0.0016	N/A	0.0023	N/A
14	0.0085	0.0616	624.71%	0.0678	697.65%
15	0.3948	0.3507	1.17%	0.3806	-3.60%
16	0.4405	0.4630	5.11%	0.4419	0.32%
17	0.0622	0.0614	.29%	0.0236	-62.06%
18	0.0299	0.0358	19.73%	0.0373	24.75%
19	0.1266	0.1314	3.79%	0.1533	21.09%
20	0.9583	0.9583	0.00%	0.9583	0.00%
21	0.6117	0.5999	.93%	0.6209	1.50%
22	1.0000	1.0000	0.00%	1.0000	0.00%
24	0.6618	0.6795	2.67%	0.7041	6.39%
25	0.3934	0.1372	-65.12%	0.0992	-74.78%
26	0.5355	0.5365	0.19%	0.4914	-8.24%
27	0.3738	0.3984	6.58%	0.3776	1.02%
28	0.5376	0.4857	-9.65%	0.3847	-28.44%

Interestingly, sometimes, as in query 6, 13 and 14, the performance is greatly enhanced

by increasing the length of dictionary words. In query 6, the only difference is how the phrase World Trade Organization is segmented. For segmenter (n=2), World Trade Organization is segmented into 世界 (world) 贸易 (trade) 组织 (organization); For segmenter (n=all possible lengths), World Trade Organization is segmented into 世界贸易组织 (World Trade Organization). Query number 14 is an interesting case to notice, since it reveals the cross ambiguity of Chinese segmentation. The difficulty lies how 爱滋病 (AIDS) should be segmented. The original sentence is 中国的爱滋病例 (China's AIDS cases), where both 爱滋病(AIDS) and 病例 (case) are in the dictionary. When the two words, 爱滋病(AIDS) and 病例 (case), are combined together, instead of repeating character 病, only one character of 病 is used. But what is the correct segmentation? Even native speakers cannot agree with each other, since both 爱滋病例 and 爱滋 病例 make sense. Segmenter (n=all possible lengths) and segmenter (n<=3) got the former, producing an average precision more than five times higher that that of segmenter (n=2) who got the latter result.

The next questions are why some relevant documents are retrieve and why some are not, and whether word segmentation is one influential factor for the results. I randomly choose some queries, where there is significant performance difference between segmenter (n=2) and segmenter (n=all possible word lengths). Take query (ID=2) as an example, the average precision for segmenter (n=2) is 0.1084, compared with 0.0279 of segmenter (n=all possible word lengths). The query segmentation result is the same for

both segmenters, that is 中共(abbreviation for Chinese Communist Party) 对于 (toward) 中国 (China) 统一 (unification) 的 (of) 立场 (standpoint). The query is translated as CCP's standpoint toward China's unification. I took a Boolean And approach, that is, to retrieve only those documents that contains all the terms in the query. After running the query, segmenter (n=2) approach produced five documents in the following order:

0 #12111 cb019002-bfw 267-86 weight=0.148121

1 #12453 cb035009-bfw-917-50 weight=0.147549

2 #7785 cb049007-bfw 537-528 weight=0.144524

3 #11988 cb013030-bbw-3542-652 weight=0.144157

4 #23315 cb039001-bfw 072 78 weight=0.141276

Segmenter (n=all possible word lengths) found only one document:

0 #12111 cb019002-bfw 267-86 weight=0.141265

First to be noted is that TREC experts mark none of the above documents as relevant. However a pattern match shows that all the query terms does appear in all the documents. As a matter of fact, whether the retrieved document, which contains all the query terms, is relevant or nor is not our utmost concern for this paper, not only because relevance judgment provided by TREC is objective and arbitrary, but also because this is more of a problem for ranking algorithm. For Chinese segmentation, we are more concerned about under which terms a document should be indexed. After a careful examination of two

documents: document cb019002-bfw 267-86, retrieved by both segmenters, and document cb035009-bfw-917-50, retrieved by segmenter (n=2) only, here is the following findings:

1. 中国 (China), 中国政府 (Chinese Government) ，中国共产党 (Chinese Communist Party) ， 中国人（Chinese）, 中国人民 (Chinese people), 中共 (abbreviation for Chinese Communist Party) ，中共中央总书记 (Secretary-General of Chinese Central Government) ，统一 (unification) ， 和平统一 (peaceful unification) ，祖国统一 (motherland unification), 两个中国 (Two Chinas), 关于（toward)， 立场 (standpoint) are in the dictionary

2. As far as the query terms concerned, both document cb019002-bfw 267-86 and document cb035009-bfw-917-50 are indexed by segmenter (n=2) under all query terms: 中国， 中共， 统一， 关于， 立场. However, document cb019002-bfw 267-86 is indexed by segmenter (n=all possible word lengths) under 中国 (China), 中国政府 (Chinese Government) ，中国共产党 (Chinese Communist Party) ， 中国人（Chinese）, 中国人民 (Chinese people), 中共 (abbreviation for Chinese Communist Party) ， 中共中央总书记 (Secretary-General of Chinese Central Government) ，统一 (unification) ， 和平统一 (peaceful unification) ，祖国统一 (motherland unification), 关于（toward）， 立场 (standpoint). Document cb035009-bfw-917-50 is indexed by segmenter (n=all possible word lengths) under:

统一 (unification)， 和平统一 (peaceful unification)，祖国统一 (motherland unification), 中国 (China), 中国共产党 (Chinese Communist Party)， 中国人（Chinese）, 两个中国 (Two Chinas), 中共中央总书记 (Secretary-General of Chinese Central Government), 关于（toward）原则立场 (principal and standpoint). We can see in Chart 3 that two words, 中共 and 立场 are missing as index terms from document cb035009-bfw-917-50 indexed by segmenter (n=all possible word lengths).

Chart 3: Summary of Term Frequencies in Two Documents with Different Segmenter

	中共	关于	中国	统一	立场
cb019002-bfw 267-86 n=2	**2**	**4**	**14**	**14**	**2**
cb019002-bfw 267-86 n=all possible lengths	**1**	**4**	**8**	**3**	**2**
cb035009-bfw-917-50 n=2	**1**	**1**	**9**	**6**	**2**
cb035009-bfw-917-50 n=all possible lengths	**0**	**1**	**3**	**1**	**0**

In one word, short word indexing can produce better retrieval results by avoiding over-specification problems of long word indexing, if short words are often used as query terms. In the above example, 中共 (abbreviation of Chinese Communist Party) instead of 中国共产党(Chinese Communist Party)，统一 (unification) instead of 祖国统一 (motherland unification) or 和平统一 (peace unification) are used in the query.

However, if long words are used in the query, such as 世界贸易组织 (World Trade Organization) in query number 6, the retrieval results can be greatly enhanced using words of all possible lengths in the dictionary. Generally the longer a word is, the less frequent it is going to appear in a document, and therefore higher weight is to be assigned to it.

We now get back to the argument of how phrases should be indexed in IR system. By providing more contextual information, long word/phrases does produce better results, if the users will use the long words/phrases in building queries. One possible solution is duplicated indexing, when one document can be indexed under both long words and short words. However it also means more storage spaces, more computing power and better algorithms to decide what to index and what not to index. It is tradeoff IR systems for all kind of languages needs to deal with.

6. Conclusion and Future Work

From our experiment, we find that less accurate segmentation will not necessarily yield worse information retrieval results. As a matter of fact, allowing two-character words only in the dictionary produced the best retrieval results in terms of precision and recall. Allowing longer words in the dictionary will lead to the missing of index words -- the problem of over-specification, which is quite common to any phrase indexing techniques. However, long-word indexing can produce better results when the long-word is used in

queries as well. Therefore, we can further enhance our segmenter by allowing redundant indexing, that is, not only index a document under the longest words, but also index it under any words that are a sub-string of this long word. In that case, we can accommodate both long-word queries and short-word queries.

Whatever is not in the dictionary will be segmented character by character. As a result our segmenter is very ineffective at identifying entity names, such as human, places. Our next step is to write a program that can automatically identify new words, including entity name, by further segment a string of single characters using statistical methods after the initial dictionary-based segmentation.

Ambiguity problems are also common in our experiments. The combined ambiguity is basically the problem of short-word and long-word indexing, but crossed ambiguity is difficult to solve, and most often produce wrong indexes. For example 活动日益猖獗 (activities are getting rampant every day) is segmented into 活动日 - 益 - 猖獗 (activity day – benefit - rampant) instead of 活动 - 日益 - 猖獗 (activity – day by day - rampant). This is a direct result of Forward Greedy Algorithm, when longer words are extracted earlier than shorter words. Even though 日 can be combined both with word before and word after it, the former is always chosen by our segmenter. This is also a problem worth investigating for future work.

Reference

Allen, J. (1995). *Natural language understanding.* Redwood City, CA: Benjamin/Cummings

Buckley, C., Singhal, A., Mitra, M. (1996). Using query zoning and correlation within SMART:TREC 5. *In Preceedings of the fifth text retrieval conference (TREC-5)*. Gaithersburg, MD, November 20-22.

Chen, Keh-Jiann & Liu, Shing-Huan. (1992) Word Identification for Mandarin Chinese Sentences. *In Proceedings of COLING-92*, pages 101 07. LOLING

Cheng, K.S., Young, G.H. & Wong, K.F. (1999) A study on word-based and integral-bit Chinese text compression algorithms. *Journal of the American Society for Information Science*, 50(3), pp. 218-228

Foo, S., & Li, H. (2002) Chinese word segmentation and its effect on information retrieval. To appear in *Information Processing & Management.* Available at http://islab.sas.ntu.edu.sg:8000/user/schubert/publications/2002/2002ipm_fmt.pdf

He, W. H. (1983) Automatic recognition of Chinese words. *Master Thesis (in Chinese)*. National Taiwan Institute of Technology.

Hoosain, R. (1991). *Psycholinguistic implications for linguistic relativity: A case study of Chinese*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Huang, X. J., & Robertson, S. E. (1997). *Okapi Chinese text retrieval experiments at TREC-6*. Available: http://trec.nist.gov/pubs/trec6/t6_proceedings.html, Maryland.

Khoo, C. S. G., Dai, Y. B. (2002) Using statistical and Contextual Information to Identify Two- and Three-Character Words in Chinese text. *Journal of American Society for Information Science and Technology*, 53(5):365-377

Lee, K. H., Ng, M. K. M, & Lu, Q. (1999). Text segmentation for Chinese spell checking.

Journal of the American Society of Information Science, 50(9): 751-759.

Leung, C.H.& Kan, W.K. (1996) Parallel Chinese word segmentation algorithm based on maximum matching. *Neural, Parallel & Science Computations*, 4(3), 291-303

Li, B. Y., Lien, S., Sun, C.F. and Sun, M. S. (1991) A maximal matching automatic Chinese word segmentation algorithm using corpus tagging for ambiguity resolution. *R.O.C Computational Linguistics Conference (ROCLING-IV)*, Taiwan, pp. 135 46.

Nie, J. Y., Chevallet, J. P., & Bruandet, M. F. (1997). *Between terms and words for European language IR and between words and bigrams for Chinese IR*. Available: http://trec.nist.gov/pubs/trec6/t6_proceedings.html, Maryland.

Palmer, D. (1997) A trainable rule-based algorithm for word segmentation. *Proceedings of the 35 th Annual Meeting of the Association for Computational Linguistics (ACL '97)*, Madrid, 1997.

Palmer, D. and Burger, J. (1997) Chinese Word Segmentation and Information Retrieval. *In AAAI Spring Symposium on Cross-Language Text and Speech Retrieval,* Electronic Working Notes.

Rijsbergen, C.J.V. (1979) *Information Retrieval.* London: Butterworths.

Smeaton, A., & Wilkinson, R. (1996). *Spanish and Chinese document retrieval in TREC-5*. Available: http://trec.nist.gov/pubs/trec5/t5_proceedings.html, Maryland.

Sproat, R., & Shih, C. (1990). A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages*, 4(4), 336–351.

Sproat, R., & Shih, C. (1996). A stochastic .nite-state word-segmentation algorithm for Chinese. *Computational Linguistics.* 22(3), 377–404.

The Number of Chinese Characters. Retrieved April 13, 2003, from Harvard University, Chinese Language Program, Dept. of East Asian Languages & Civilizations website: http://www.fas.harvard.edu/~clp/China/quantity.htm

Tsai, C.H. (2002) *A Review of Chinese Word Lists Accessible on the Internet* <http://www.geocities.com/hao510/wordlist/>

Tsai, C. H., McConkie, G. W., & Zheng, X. (1998, November). Lexical parsing by Chinese readers. *Poster session presented at the Advanced Study Institute on Advances in Theoretical Issues and Cognitive Neuroscience Research of the Chinese Language*, University of Hong Kong.

Wang, Y. (1985). On Chinese word division by computer. *Journal of Applied Sciences.* 3, 114 18 (in Chinese)

Wang, Y., & Xiao, W. (1984). Automatic compilation of Chinese key-word-in-title rotation index and automatic word extraction. *Journal of Nanjing University (National Science Edition)*. 1, 39-44 (in Chinese).

Wu, Z. M. & Tseng, G. (1995) ACTS: an automatic Chinese text segmentation system for full text retrieval. *Journal of American Society for Information Science.* 46(2):83-96

Yao, T. S., Zhang, G.P. & Wu, Y. M. (1990) A rule-based Chinese automatic segmentation system. *Journal of Chinese Information Processing.* 4(1):37-43

Yeh, C. L. & Lee, H. J. (1991) Rule-based word identification for Mandarin Chinese sentences – A unification approach. *Computer processing of Chinese and Oriental Languages*. 5(2): 97 18

Yu, J.S. & Yu, S. W. Some Problem of Chinese Segmentation. <http://icl.pku.edu.cn/yujs/papers/pdf/onseg.pdf> (March 22, 2003)

Appendix A

Sample TREC document:

<DOC>
<DOCID> CB019002.BFW (1267) </DOCID>
<DOCNO> CB019002-BFW-1267-86 </DOCNO>
<DATE> 1995-02-02 15:58:30 (8) </DATE>
<TEXT>
<headline> 中国和平统一促进会高度评价江泽民关于台湾问题的重要讲话 </headline>

<s> 新华社北京二月二日电（记者方瑾 陈建山）中国和平统一促进会部分在京理事今天在此间举行的座谈会上指出，中共中央总书记、国家主席江泽民关于台湾问题的重要讲话，对于进一步推动两岸关系向前发展、加快祖国和平统一的进程意义深远。 </s>
<s> 统促会今后将更广泛地联系海内外各界人士及相关团体，为推动祖国统一大业早日实现作出贡献。 </s>

<s> 全国政协副主席、统促会执行会长钱伟长主持了今天的座谈会。 </s>
<s> 统促会人士近三十人出席了座谈会。 </s>

<s> 与会者认为，江泽民总书记在讲话中提出的关于促进祖国统一大业早日完成的一系列建议和主张，表明了中国共产党和中国政府实现祖国和平统一的目标是坚定不移的，显示了中国共产党、中国政府对两岸和平统一的诚意。 </s>

<s> 全国政协副主席、统促会会长万国权在发言中说，江总书记的讲话原则性很强，阐明了我们主张“和平统一、一国两制”的立场，并重申坚决反对任何搞“一中一台”、“台独”的图谋和行动。 </s>
<s> 他说，江泽民在讲话中主张就“正式结束两岸敌对状态、逐步实现和平统一”进行谈判，这一重要宣示，在两岸关系日益向前发展的今天，有着很强的现实意义。 </s>

<s> 全国人大常委会副委员长、统促会会长程思远说，世界上只有一个中国，台湾是中国的一部分，分裂是民族的不幸，是海峡两岸的中国人都不愿看到的。 </s>
<s> 凡是中华民族的子孙，都希望中国统一，闹分裂不但违背中华民族根本利益，也是不得人心的。 </s>
<s> 江泽民总书记在讲话中主张通过谈判接触来解决两岸的分歧，以和平的方式实现国家的统一，讲得明白恳切，合情合理。 </s>

<s> 机不可失、事不宜迟，台湾当局对这些主张和建议，应该慎重考虑。 </s>

<s> 统促会理事、中国社科院台湾研究所研究员李家泉认为，“坚持统一、反对分裂”，是江泽民总书记全篇讲话的主旨。 </s>
<s> 他说，在如何处理台湾问题、发展两岸关系上有两种做法：一是在和平统一的方针下坚持一个中国、两制并存、高度自治、和平谈判，二是坚持分裂，搞“分裂分治”、以拖求变。 </s>
<s> 两种做法必将出现两种前途：前者是两岸和解、携手合作、共建祖国、共振中华；后者是拒绝和解、同室操戈、骨肉相残、亲痛仇快。 </s>
<s> 他希望台湾当局以民族大义为重，尽早作出明智选择。 </s>

<s> 统促会理事、中国人民大学教授方生指出，江泽民总书记提出的一系列建议中，很重要的一条是强调要大力发展两岸经济交流与合作，以利两岸经济共同繁荣，造福整个中华民族。</s>
<s> 江泽民关于“不以政治分歧去影响、干扰两岸经济合作”的主张，体现了中共积极发展两岸关系的一贯立场，反映了两岸人民的共同愿望和切身利益。 </s>
<s> 这也是在两岸政治歧见一时得不到解决的情况下，加强两岸经济合作的唯一可行的正确方针。 </s>

<s> 全国人大常委会副委员长、统促会会长王光英，统促会理事、中国社科院政治学研究所所长吴大英，统促会常务理事、全国政协常委冯理达等人在发言中表示，在新的一年里，统促会应大力宣传，认真学习、贯彻江泽民总书记的重要讲话，更广泛地联合海内外各界人士，发展海峡两岸的经济、文化、科技等各方面的交流与合作，促进两岸直接“三通”，为推动祖国统一大业的早日实现作出更大贡献。 </s>
<s> （完） </s>

</TEXT>

Appendix B

Chinese queries for TREC-5 Chinese Collection:

CH1 美国决定将中国大陆的人权状况与其是否给予中共最惠国待遇分离
CH2 中共对于中国统一的立场
CH3 中共核电站之营运情况
CH4 中国大陆新发现的油田
CH5 中国有关知识产权的立法与政策以及执法情况
CH6 国际社会对中共加入世界贸易组织所给予之支持
CH7 中国大陆与台湾对南海诸岛的立场
CH8 地震在日本造成的损害与伤亡数据
CH9 中国毒品问题
CH10 新疆的边境贸易
CH11 联合国驻波斯尼亚维和部队
CH12 世界妇女大会
CH13 中国争取举办西元 2000 年奥运
CH14 中国的爱滋病例
CH15 联合国维和部队如何帮助海地恢复民主制度
CH16 联合国对伊拉克经济制裁的辩论
CH17 中国对亚太经济合作组织的期望
CH18 中东和平会议
CH19 希望工程
CH20 越战失踪美军
CH21 香港总督彭定康在香港回归中国一事上所扮演的角色
CH22 世界各地感染疟疾的情况
CH23 苏联在海湾战争中如何担任调停的角色
CH24 对取消向波黑穆斯林武器禁运的反应
CH25 中国对熊猫的保护
CH26 中国森林火灾的防范措施
CH27 中国在机器人方面的研制
CH28 移动电话在中国的成长

Appendix C

Query segmentation using n-character words segmenter where n=2

1 美国 决定 将 中国 大陆 的人 权 状况 与其 是否 给予 中共 最 惠 国 待遇 分离
2 中共 对于 中国 统一 的 立场
3 中共 核电 站 之 营运 情况
4 中国 大陆 新 发现 的 油田
5 中国 有关 知识 产权 的 立法 与 政策 以及 执法 情况
6 国际 社会 对 中共 加入 世界 贸易 组织 所 给予 之 支持
7 中国 大陆 与 台湾 对 南海 诸岛 的 立场
8 地震 在 日本 造成 的 损害 与 伤亡 数据
9 中国 毒品 问题
10 新疆 的 边境 贸易
11 联合 国 驻波 斯 尼亚 维和 部队
12 世界 妇女 大会
13 中国 争取 举办 西元 年 奥运
14 中国 的 爱滋 病例
15 联合 国 维和 部队 如何 帮助 海地 恢复 民主 制度
16 联合 国 对 伊 拉 克 经济 制裁 的 辩论
17 中国 对 亚太 经济 合作 组织 的 期望
18 中东 和平 会议
19 希望 工程
20 越战 失踪 美军
21 香港 总督 彭 定 康 在 香港 回归 中国 一事 上 所 扮演 的 角色
22 世界 各地 感染 疟疾 的 情况
23 苏联 在 海湾 战争 中 如何 担任 调停 的 角色
24 对 取消 向 波黑 穆 斯 林 武器 禁运 的 反应
25 中国 对 熊猫 的 保护
26 中国 森林 火灾 的 防范 措施
27 中国 在 机器 人 方面 的 研制
28 移动 电话 在 中国 的 成长

Query segmentation using n-character words segmenter where n<=3

1 美国 决定 将 中国 大陆 的人 权 状况 与其 是否 给予 中共 最惠国 待遇 分离
2 中共 对于 中国 统一 的 立场
3 中共 核电站 之 营运 情况
4 中国 大陆 新发现 的 油田
5 中国 有关 知识 产权 的 立法 与 政策 以及 执法 情况
6 国际 社会 对 中共 加入 世界 贸易 组织 所 给予 之 支持
7 中国 大陆 与 台湾 对 南海 诸岛 的 立场
8 地震 在 日本 造成 的 损害 与 伤亡 数据
9 中国 毒品 问题
10 新疆 的 边境 贸易
11 联合国 驻波 斯 尼亚 维和 部队
12 世界 妇女 大会
13 中国 争取 举办 西元 年 奥运
14 中国 的 爱滋病 例
15 联合国 维和 部队 如何 帮助 海地 恢复 民主 制度
16 联合国 对 伊拉克 经济 制裁 的 辩论
17 中国 对 亚太 经济 合作 组织 的 期望
18 中东 和平 会议
19 希望 工程
20 越战 失踪 美军
21 香港 总督 彭 定 康 在 香港 回归 中国 一事 上 所 扮演 的 角色
22 世界 各地 感染 疟疾 的 情况
23 苏联 在 海湾 战争 中 如何 担任 调停 的 角色
24 对 取消 向 波黑 穆斯林 武器 禁运 的 反应
25 中国 对 熊猫 的 保护
26 中国 森林 火灾 的 防范 措施
27 中国 在 机器人 方面 的 研制
28 移动 电话 在 中国 的 成长

Query segmentation using n-character words segmenter where n<=4

1 美国 决定 将 中国大陆 的人 权 状况 与其 是否 给予 中共 最惠国 待遇 分离
2 中共 对于 中国 统一 的 立场
3 中共 核电站 之 营运 情况
4 中国大陆 新发现 的 油田
5 中国 有关 知识产权 的 立法 与 政策 以及 执法 情况
6 国际 社会 对 中共 加入 世界贸易 组织 所 给予 之 支持
7 中国大陆 与 台湾 对 南海诸岛 的 立场
8 地震 在 日本 造成 的 损害 与 伤亡 数据
9 中国 毒品 问题
10 新疆 的 边境贸易
11 联合国 驻波 斯 尼亚 维和 部队
12 世界 妇女 大会
13 中国 争取 举办 西元 年 奥运
14 中国 的 爱滋病 例
15 联合国 维和 部队 如何 帮助 海地 恢复 民主 制度
16 联合国 对 伊拉克 经济 制裁 的 辩论
17 中国 对 亚太 经济 合作 组织 的 期望
18 中东 和平 会议
19 希望 工程
20 越战 失踪 美军
21 香港 总督 彭 定 康 在 香港 回归中国 一事 上 所 扮演 的 角色
22 世界各地 感染 疟疾 的 情况
23 苏联 在 海湾战争 中 如何 担任 调停 的 角色
24 对 取消 向 波黑 穆斯林 武器 禁运 的 反应
25 中国 对 熊猫 的 保护
26 中国 森林 火灾 的 防范 措施
27 中国 在 机器人 方面 的 研制
28 移动电话 在 中国 的 成长

Query segmentation using n-character words segmenter where n<=5

1 美国 决定 将 中国大陆 的人 权 状况 与其 是否 给予 中共 最惠国待遇 分离
2 中共 对于 中国 统一 的 立场
3 中共 核电站 之 营运 情况
4 中国大陆 新发现 的 油田
5 中国 有关 知识产权 的 立法 与 政策 以及 执法 情况
6 国际 社会 对 中共 加入 世界贸易 组织 所 给予 之 支持
7 中国大陆 与 台湾 对 南海诸岛 的 立场
8 地震 在 日本 造成 的 损害 与 伤亡 数据
9 中国 毒品 问题
10 新疆 的 边境贸易
11 联合国 驻波 斯 尼亚 维和 部队
12 世界 妇女 大会
13 中国 争取 举办 西元 年 奥运
14 中国 的 爱滋病 例
15 联合国 维和 部队 如何 帮助 海地 恢复 民主 制度
16 联合国 对 伊拉克 经济 制裁 的 辩论
17 中国 对 亚太 经济 合作 组织 的 期望
18 中东 和平 会议
19 希望 工程
20 越战 失踪 美军
21 香港 总督 彭 定 康 在 香港 回归中国 一事 上 所 扮演 的 角色
22 世界各地 感染 疟疾 的 情况
23 苏联 在 海湾战争 中 如何 担任 调停 的 角色
24 对 取消 向 波黑 穆斯林 武器 禁运 的 反应
25 中国 对 熊猫 的 保护
26 中国 森林 火灾 的 防范 措施
27 中国 在 机器人 方面 的 研制
28 移动电话 在 中国 的 成长

Query segmentation using n-character words segmenter where 2<=n<=22

1 美国 决定 将 中国大陆 的人 权 状况 与其 是否 给予 中共 最惠国待遇 分离
2 中共 对于 中国 统一 的 立场
3 中共 核电站 之 营运 情况
4 中国大陆 新发现 的 油田
5 中国 有关 知识产权 的 立法 与 政策 以及 执法 情况
6 国际 社会 对 中共 加入 世界贸易组织 所 给予 之 支持
7 中国大陆 与 台湾 对 南海诸岛 的 立场
8 地震 在 日本 造成 的 损害 与 伤亡 数据
9 中国 毒品 问题
10 新疆 的 边境贸易
11 联合国 驻波 斯 尼亚 维和 部队
12 世界 妇女 大会
13 中国 争取 举办 西元 年 奥运
14 中国 的 爱滋病 例
15 联合国 维和 部队 如何 帮助 海地 恢复 民主 制度
16 联合国 对 伊拉克 经济 制裁 的 辩论
17 中国 对 亚太 经济 合作 组织 的 期望
18 中东 和平 会议
19 希望 工程
20 越战 失踪 美军
21 香港 总督 彭 定 康 在 香港 回归中国 一事 上 所 扮演 的 角色
22 世界各地 感染 疟疾 的 情况
23 苏联 在 海湾战争 中 如何 担任 调停 的 角色
24 对 取消 向 波黑 穆斯林 武器 禁运 的 反应
25 中国 对 熊猫 的 保护
26 中国 森林 火灾 的 防范 措施
27 中国 在 机器人 方面 的 研制
28 移动电话 在 中国 的 成长